

Stabilizing Value Iteration with and without Approximation Errors

Ali Heydari¹

Abstract—Adaptive optimal control using value iteration (VI) initiated from a stabilizing policy is theoretically analyzed in various aspects including the continuity of the result, the stability of the system operated using any *single/constant* resulting control policy, the stability of the system operated using the *evolving/time-varying* control policy, the convergence of the algorithm, and the optimality of the limit function. Afterwards, the effect of presence of approximation errors in the involved function approximation processes is incorporated and another set of results for boundedness of the approximate VI as well as stability of the system operated under the results for both cases of applying a single policy or an evolving policy are derived. A feature of the presented results is providing estimations of the *region of attraction* so that if the initial condition is within the region, the whole trajectory will remain inside it and hence, the function approximation results will be reliable.

I. INTRODUCTION

Intelligent control using adaptive/approximate dynamic programming (ADP), sometimes referred to by reinforcement learning (RL) or neuro-dynamic programming (NDP), is a set of powerful tools for obtaining approximate solutions to difficult and mathematically intractable problems which seek optimum while sometimes even no knowledge of the system model/dynamics is available. The dramatic potential of the tools in practice has attracted many researchers within the last few decades, [1]- [13]. The multitude of appeared papers and success stories on applications of ADP to different problems, however, has intensified the need for firm mathematical analyses for guaranteeing the convergence of the learning processes and the stability of the results.

Besides the classifications of heuristic dynamic programming (HDP), dual heuristic programming (DHP), etc. [7], which are in terms of the variables subject to approximation and their dependencies, the learning algorithms are typically based on either value iteration (VI) or policy iteration (PI), [3], [14]. These algorithms are well investigated both by computer scientists for machine learning [3] and by control scientists for feedback control of dynamical systems [14]. PI, despite having a higher computational load due to a ‘full backup’ as opposed to a ‘partial backup’ in VI [14], has the advantage that the control under evolution remains stabilizing, [15]. Hence, PI seems more suitable for online implementation, i.e., adapting the control ‘on the fly’. However, the requirement that PI needs to start with an stabilizing initial control is one of its drawbacks. VI, on the other hand, does not require an stabilizing initial control and can be initiated arbitrarily. But, the closed loop system is not guaranteed to be stable during its learning process, if implemented online.

Considering optimal control of discrete-time problems with continuous state and action spaces and undiscounted cost functions using VI, which is the subject of this work, the convergence proof for linear systems was analyzed in [16], [17]. As for nonlinear systems, the convergence was established by different researchers including [18] (adapted in [19]), [10], and [20] through different approaches. All these convergence analyses are based on the assumption of perfect function reconstruction, i.e., no error in the function approximation. While this assumption plays a major rule in deriving the results, it restricts their practical use severely, because, the approximation errors exist almost in every application when the system is nonlinear or when the cost function terms are non-quadratic and nonlinear. What makes their presence potentially problematic is the fact that the errors *propagate* throughout the iterations, hence, regardless of how small they are, a phenomenon similar to *resonance* might happen which could lead to the complete unreliability of the results.

Analyzing VI under the presence of approximation errors, i.e. *approximate VI (AVI)*, is an open research problem with a few published results, including [4], [21], [22], [23], [12], to the best of the knowledge of the author. Refs. [4], [21], [22], [23] investigated problems with *discounted* cost functions and the results are solely valid for such problems, prevalent in computer science. As a matter of fact, the ‘forgetting’ nature of discounted problems is the backbone of the developments of the error bounds and if the discount factor approaches one, as in typical infinite-horizon optimal control problems, the bounds go to infinity. Hence, the results do not cover this case. On the other hand, the interesting results in [12] provide some error analyses but with assumptions which are more restrictive and not easily verifiable, compared with this study. For example, the approximation error between the exact and approximate functions, respectively denoted with $V(\cdot)$ and $\hat{V}(\cdot)$, should be possible to be written in the *multiplicative* form of $\hat{V}(x) \leq \sigma V(x)$ for some positive constant $\sigma, \forall x$, instead of an *additive* form of $\hat{V}(x) = V(x) + \epsilon(x)$, for some real valued function $\epsilon(\cdot)$. Moreover, the boundedness results are conditional upon σ being upper bounded with a term including a parameter which corresponds to the *optimal* value function.

Based on this background, besides the stability issue during the online learning stage using VI, rigorous theoretical analyses of the consequences of the errors on the results are of great interest to the ADP researchers and practitioners. The reasons are the scarcity of the available studies on AVI, the prevalence of approximation errors, and the great potential of the tool in (approximately) solving optimal control problems in practice.

The contributions of this study are multiple. Initially, it is proved that VI also will be stabilizing for online control if,

¹Assistant Professor of Mechanical Engineering, South Dakota School of Mines and Technology, Rapid City, SD 57701, email: ali.heydari@sdsmt.edu.

similar to PI, it is started using an initial stabilizing control. Afterwards, it is shown that the start from an initial stabilizing control leads to an initial value function that does not satisfy the necessary conditions for any of the cited convergence proofs of VI. Establishing this convergence (to the optimal solution) is another contribution of this work. These results may not look substantially different from what the ADP community assumes to hold intuitively or has already established, [24]. What makes the abovementioned two results different is having two characteristics. The theoretical analyses in this study are simple and straight forward, both for optimality and stability analyses, compared to the existing developments in the literature. Another feature is providing rigorous mathematical bases for the analyses. As an example, the use of a value function as a Lyapunov function for stability analysis requires proof of continuity of the value function, [25]. The firm proof of this continuity, presented in this work, is not as straight forward as it looks. The factors leading to the difficulty are the presence of the *argmin* operator in calculation of the control at each iteration of VI, which may potentially lead to a discontinuity in the control policy, and also the concern of *pointwise* versus *uniform* convergence, for concluding the continuity of the limit function from the continuity of the elements of a converging sequence of functions, [26], [19].

Another contribution of this work is addressing the legitimate concern that any ADP result is valid only when the state trajectory remains within the domain for which the controller is trained. This concern is resolved through establishing an *estimation of the region of attraction* (EROA) [25] for the controller, in this work, so that as long as the system's initial condition is within the region, it is guaranteed that the entire trajectory remains in the region. Hence, the controller will remain valid and usable.

As the reader delves into the problem, it is discussed that the provided stability proof, whose main idea is not much different from [24], assumes applying a fixed (time-invariant) control policy on the system. But, this is rarely the case in online learning, since, as the learning proceeds, the control policy evolves, hence, the applied control policy is time-varying. Therefore, another set of stability results for the time-varying and evolving control policy is developed with some ideas for establishing its respective EROA, as another contribution of this work.

After providing the detailed analysis of the VI which is initiated with an admissible guess, called Stabilizing VI throughout the paper, the case of presence of the approximation errors is investigated, leading to a new set of contributions which are of greater interests. They include boundedness/convergence analysis of the AVI initiated with an admissible guess, stability and EROA analysis for the case of applying a fixed control policy, and stability and EROA analysis for the case of applying an evolving control policy. These theoretical analyses are the most important contributions of this work, as the assumptions leading to the results are verifiable and more straight forward, compared with the available studies. Finally, interested readers are referred to [27] for some recent developments of this author on analyzing the effect of the approximation errors in regular value iteration, i.e., the approximate VI which is

initiated arbitrarily.¹

The rest of this paper is organized as follows. The problem is formulated in Section II and the ADP-based solutions are revisited in Section III. Section IV presents the theoretical analyses on exact VI. The respective analyses for AVI are presented in section V. Finally, concluding remarks are given in Section VI.

II. PROBLEM FORMULATION

Let the system subject to control be given by discrete-time nonlinear dynamics

$$x_{k+1} = f(x_k, u_k), k \in \mathbb{N}, \quad (1)$$

where $f : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}^n$ is a Lipschitz continuous function versus its both inputs, i.e., the state and control vectors, x and u , respectively, with $f(0, 0) = 0$. The set of non-negative integers is denoted with \mathbb{N} , and positive integers n and m denote the dimensions of the continuous state and control spaces. Finally, sub-index k represents the discrete time index. The performance index is given by

$$J = \sum_{k=0}^{\infty} U(x_k, u_k), \quad (2)$$

where *utility function* $U(\cdot, \cdot)$ is of form $U(x_k, u_k) := Q(x_k) + u_k^T R u_k$ for a continuous and positive semi-definite function $Q : \mathbb{R}^n \rightarrow \mathbb{R}_+$ and a positive definite $m \times m$ real matrix R . Set \mathbb{R}_+ denotes the non-negative reals. Starting with any initial feedback *control policy* given by $h : \mathbb{R}^n \rightarrow \mathbb{R}^m$ for control calculation, i.e., $u_k = h(x_k)$, the problem is *updating/adapting* the control policy such that cost function (2) is minimized. The control policy which leads to such a characteristic is called *optimal control policy*, denoted with $h^*(\cdot)$.

Definition 1. A control policy is defined to be *asymptotically stabilizing within a domain* if $\lim_{k \rightarrow \infty} x_k = 0$ using this control policy, for every initial state within the domain, [25].

Definition 2. A control policy $h(\cdot)$ is defined to be *admissible within a compact set* if a) it is a Lipschitz continuous function of x in the set with $h(0) = 0$, b) it asymptotically stabilizes the system within the set, and c) there exists a continuous positive definite function $W : \mathbb{R}^n \rightarrow \mathbb{R}_+$ that puts an upper bound on the respective 'cost-to-go' or 'value function', denoted with $V_h : \mathbb{R}^n \rightarrow \mathbb{R}_+$ and defined by

$$V_h(x_0) = \sum_{k=0}^{\infty} U(x_k^h, h(x_k^h)), \quad (3)$$

i.e., $V_h(x) \leq W(x), \forall x \in \Omega$. In Eq. (3) one has $x_k^h := f(x_{k-1}^h, h(x_{k-1}^h)), \forall k \in \mathbb{N} - \{0\}$, and $x_0^h := x_0$. In other words, x_k^h denotes the k th element on the state trajectory/history initiated from x_0 and propagated using control policy $h(\cdot)$.

The main difference between the defined admissibility and the ones typically utilized in the ADP/RL literature, including [10], is the assumption of upper boundedness of the value function by a continuous (positive definite) function. This

¹It must be added that the current version of this paper has overlaps with the first version of [27] on Lemma 5 and Theorem 6.

condition is trivially satisfied if the value function itself is continuous, i.e., through selecting $W(\cdot) = V_h(\cdot)$. However, instead of assuming continuity of the value function, the milder condition of being upper bounded by such a function is assumed. Note that the continuity of the value function is required for *uniform* approximation of the function using parametric function approximators, [28], [29]. In this study, it will be shown that the upper boundedness will lead to the desired continuity. Finally, it should be noted that continuous functions are bounded in a compact set [26], hence, the upper boundedness of the value function by the continuous function $W(\cdot)$ leads to the boundedness of the respective value function. This is an essential requirement for an admissible control, as a mere asymptotically stabilizing control policy may lead to an unbounded value function.

Assumption 1. *There exists at least one admissible control policy for the given system within a connected and compact set $\Omega \subset \mathbb{R}^n$ containing the origin.*

Assumption 2. *The intersection of the set of n -vectors x at which $U(x, 0) = 0$ with the invariant set of $f(\cdot, 0)$ only contains the origin.*

Assumption 1 guarantees that there is no state vector in Ω for which the value function associated with the *optimal* control policy is infinite. Assumption 2 assures that there is no set of states (besides the set containing only the origin) in which the state trajectory can *hide* forever, in the sense that the utility function evaluated at those states is zero without convergence of the states to the origin. Note that, if such a set exists, then starting from an initial state within the set, the optimal solution would be $u_k = 0, \forall k$.

III. ADP-BASED SOLUTIONS

Based on Eq. (3), it can be seen that the value function satisfies the recursive relation given by

$$V_h(x) = U(x, h(x)) + V_h(f(x, h(x))), \forall x \in \mathbb{R}^n. \quad (4)$$

Defining the *optimal value function*, as the value function associated with the optimal control policy and denoting it with $V^*(\cdot)$, the Bellman equation [30], given below, provides the solution to the problem

$$h^*(x) = \operatorname{argmin}_{u \in \mathbb{R}^m} (U(x, u) + V^*(f(x, u))), \quad (5)$$

$$V^*(x) = \min_{u \in \mathbb{R}^m} (U(x, u) + V^*(f(x, u))). \quad (6)$$

Due to the *curse of dimensionality* [30], however, the proposed solution is mathematically impracticable for general nonlinear systems. ADP utilizes the idea of *approximating* the optimal value function, using either look-up tables or function approximators, e.g., neural networks (NNs), for remedying the problem. The value function approximator is typically called the *critic* in the ADP/RL literature. The approximation is performed over a *compact* and *connected* set containing the origin, called the *domain of interest*. This domain, denoted with Ω , has to be selected based on the specific problem at hand and it should be noted that the ADP based results are valid only if the entire state trajectory initiated from the

initial state vector remains within the domain for which the value function is approximated. The optimal value function approximation process is typically done through PI or VI. In PI, one starts with an initial admissible control policy, denoted with $h^0(\cdot)$, and iterates through the *policy evaluation equation* given by

$$V^i(x) = U(x, h^i(x)) + V^i(f(x, h^i(x))), \forall x \in \Omega, \quad (7)$$

and the *policy update equation* given by

$$h^{i+1}(x) = \operatorname{argmin}_{u \in \mathbb{R}^m} (U(x, u) + V^i(f(x, u))), \forall x \in \Omega, \quad (8)$$

for $i = 0, 1, \dots$ until the parameters converge. In other words, starting from policy $h^0(\cdot)$, pointwise values for approximating value function $V^0(\cdot)$ can be calculated using (7) and once $V^0(\cdot)$ is used in (8) pointwise values for approximating $h^1(\cdot)$ can be calculated, and so on.

On the other hand, in VI, the iterative learning starts with an initial guess $V^0(\cdot)$ and iterates through the *policy update equation* given by

$$h^i(x) = \operatorname{argmin}_{u \in \mathbb{R}^m} (U(x, u) + V^i(f(x, u))), \forall x \in \Omega, \quad (9)$$

and the *value update equation*

$$V^{i+1}(x) = U(x, h^i(x)) + V^i(f(x, h^i(x))), \forall x \in \Omega, \quad (10)$$

or equivalently

$$V^{i+1}(x) = \min_{u \in \mathbb{R}^m} (U(x, u) + V^i(f(x, u))), \forall x \in \Omega, \quad (11)$$

for $i = 0, 1, \dots$ until the iterations converge.

IV. ANALYSIS OF STABILIZING VALUE ITERATION

PI requires an admissible policy $h^0(\cdot)$ to start the process with, otherwise, there may not exist a bounded (and continuous) function $V^0(\cdot)$ which satisfies Eq. (7) for $i = 0$. This can be observed by realizing that $V^0(\cdot)$ is actually the value function associated with policy $h^0(\cdot)$. To confirm this one may compare Eq. (7) with Eqs. (4) and (3).

An important feature of PI is the fact that $V^i(\cdot)$ s remain Lyapunov functions for the closed loop system, hence, each respective control policy will be stabilizing. This can be confirmed by noting that from Eq. (7) one has

$$\begin{aligned} \Delta V^i(x) &:= V^i(f(x, h^i(x))) - V^i(x) = \\ &\quad - U(x, h^i(x)) \leq 0, \forall x \in \Omega. \end{aligned} \quad (12)$$

This feature leads to its suitability for *online* implementation, i.e., for *adaptive* optimal control. It is an important point that *online* learning has the advantage of not requiring perfect knowledge of the internal dynamics of the system [10]. However, stability of the system operated using the evolving control is of critical importance.

The VI scheme, however, can be arbitrarily initiated using any $V^0(\cdot)$. The convergence to the optimal solution is proved for $V^0(x) = 0, \forall x$, in [10] and for any smooth $V^0(\cdot)$ which satisfies $0 \leq V^0(x) \leq U(x, 0), \forall x$, in [20]. Moreover, the convergence is proved in [18] for $\alpha V^*(x) \leq V^0(x) \leq V^*(x), \alpha \in [0, 1], \forall x$, assuming there exists a finite β , independent of x , such that $V^*(f(x, u)) \leq \beta U(x, u), \forall x$.

VI is more interesting for control of nonlinear systems as the convergence to optimal solution is guaranteed without requiring an initial admissible control, and also, instead of solving equation (7), known as a ‘full backup’, the ‘partial backup’ given by the simple recursion (10) is needed, [14]. However, these advantages come with the disadvantage that the ‘immature’ control, i.e., $h^i(\cdot)$ s before the convergence of the solution, are not guaranteed to be stabilizing. This issue is generally considered as a shortcoming of VI as opposed to PI in the ADP literature, see [15] as an example. In this study, it will be shown that the credit for stabilizing feature of $h^i(\cdot)$ s in PI is due to the initial admissible control and if VI also is initiated with such a control policy, the resulting control policies under iterations will be stabilizing.

Let the initial guess, $V^0(\cdot)$, to be used in VI, be given by the value function of an admissible control policy. For the sake of brevity, such a VI is called *stabilizing value iteration*, throughout the paper, as defined in the next definition.

Definition 3. *The value iteration scheme given by recursive relation (11) which is initiated using the value function of an admissible control policy is called stabilizing value iteration.*

Denoting the initial admissible policy with $h^{-1}(\cdot)$ (for notational compatibility), its value function, denoted with $V^0(\cdot)$, can be calculated through solving

$$V^0(x) = U(x, h^{-1}(x)) + V^0(f(x, h^{-1}(x))), \forall x \in \Omega, \quad (13)$$

per (4). Utilizing this $V^0(\cdot)$ in the VI as the initial guess, the stability of $h^i(\cdot)$ s can be guaranteed, as proved here, which requires some theoretical results given next.

Lemma 1. *Sequence of functions $\{V^j(x)\}_{j=0}^\infty := \{V^0(x), V^1(x), \dots\}$ resulting from stabilizing value iteration is a pointwise non-increasing sequence.*

Proof: The proof is done by induction. Considering (13), which gives $V^0(\cdot)$, and (11), which (for $i = 0$) gives $V^1(\cdot)$, one has

$$V^1(x) \leq V^0(x), \forall x \in \Omega, \quad (14)$$

because $V^1(\cdot)$ is the result of minimization of the right hand side of (13) instead of being resulted from a given $h^{-1}(\cdot)$. Now, assume that for some i , we have

$$V^i(x) \leq V^{i-1}(x), \forall x \in \Omega. \quad (15)$$

Define $\mathcal{V}(\cdot)$ as

$$\mathcal{V}(x) := U(x, h^{i-1}(x)) + V^i(f(x, h^{i-1}(x))), \forall x \in \Omega. \quad (16)$$

Comparing (16) with (11), one has

$$V^{i+1}(x) \leq \mathcal{V}(x), \forall x \in \Omega, \quad (17)$$

because $V^{i+1}(\cdot)$ is the result of minimization of the right hand side of (16). Moreover, $V^i(\cdot)$ is given by

$$V^i(x) = U(x, h^{i-1}(x)) + V^{i-1}(f(x, h^{i-1}(x))), \forall x \in \Omega, \quad (18)$$

based on (10). Hence,

$$\mathcal{V}(x) \leq V^i(x), \forall x \in \Omega, \quad (19)$$

because of (15). Considering (17) and (19), one has

$$V^{i+1}(x) \leq V^i(x), \forall x \in \Omega, \quad (20)$$

which together with (15), proves the lemma, by induction. \square

Before proceeding to the stability theorem, it is needed to analyze the continuity of each $V^i(\cdot)$ within Ω . Note that even though functions $f(\cdot, \cdot)$ and $U(\cdot, \cdot)$ are assumed to be continuous versus all the inputs, the existence of *argmin* operator in Eq. (9) may lead to discontinuity in function $h^i(\cdot)$, which may then lead to a discontinuous $V^{i+1}(\cdot)$ in Eq. (10). Besides the *argmin* issue, does the continuity of $h^{-1}(\cdot)$ lead to the continuity of its value function, $V^0(\cdot)$? Note that $V^0(\cdot)$ is going to be the initial guess in the stabilizing VI, hence, its continuity matters. Moreover, how does one find the value function of a given $h^{-1}(\cdot)$? The answer to these questions are given first. Let $V(\cdot) \in \mathcal{C}(x)$ (respectively, $V(\cdot) \in \mathcal{C}(\Omega)$) denote that function $V(\cdot)$ is continuous at point x (respectively, within Ω).

Lemma 2. *If $h(\cdot)$ is an admissible control policy within Ω , then selecting any $V_h^0(\cdot) \in \mathcal{C}(\Omega)$ which satisfies $0 \leq V_h^0(x) \leq U(x, 0), \forall x \in \Omega$, the iterations given by*

$$V_h^{j+1}(x) = U(x, h(x)) + V_h^j(f(x, h(x))), \forall x \in \Omega. \quad (21)$$

converges monotonically to the value function of $h(\cdot)$.

Proof: It is known that value function $V_h(\cdot)$ is the *fixed point* of iterations indexed by j and given by (21), [14]. The reason is Eq. (21) is equivalent of

$$V_h^j(x_0) = V_h^0(x_j^h) + \sum_{k=0}^{j-1} U(x_k^h, h(x_k^h)). \quad (22)$$

Comparing (22) with (3) and considering $0 \leq V_h^0(x) \leq U(x, 0)$ one has $0 \leq V_h^j(x) \leq V_h(x)$. Therefore, sequence $\{V_h^j(x)\}_{j=0}^\infty$ is upper bounded by $V_h(x)$. The limit function $V_h^\infty(\cdot)$ is equal to $V_h(\cdot)$, since, the admissibility of $h(\cdot)$ leads to $x_j^h \rightarrow 0$, and hence, $V_h^0(x_j^h) \rightarrow 0$ as $j \rightarrow \infty$, due to $0 \leq V_h^0(x) \leq U(x, 0)$. Hence, (22) converges to (3) as $j \rightarrow \infty$. This proves *pointwise* convergence of the sequence to $V_h(\cdot)$.

As for monotonicity of this convergence, not that for any arbitrary positive integers j_1 and j_2 , if $j_1 \leq j_2$, then

$$V_h^{j_1}(x_0) - V_h^{j_2}(x_0) = V_h^0(x_{j_1}^h) - V_h^0(x_{j_2}^h) - \sum_{k=j_1}^{j_2-1} U(x_k^h, h(x_k^h)) \leq 0, \quad (23)$$

since $0 \leq V_h^0(x_{j_1}^h) \leq U(x_{j_1}^h, 0) \leq U(x_{j_1}^h, h(x_{j_1}^h))$, and the last term in the foregoing inequality is only one of the non-negative terms in the summation in the right hand side of (23). Therefore, sequence of functions $\{V_h^j(x)\}_{j=0}^\infty$ is pointwise non-decreasing. \square

While the foregoing lemma helps in finding the value function of a given control policy, it does not prove the possible continuity of the result. The reason is, even though $V_h^j(\cdot)$ is continuous for any $j < \infty$, as a finite sum of continuous functions given in (22), the limit function may not be continuous, as only the *pointwise* convergence is proved. An idea for proof of uniform continuity can be adapted from [19], which is based on [18], and hence, requires $V_h(f(x, u)) \leq \beta U(x, u)$ to hold uniformly for a constant β . However, the foregoing condition restricts the generality of the result. Besides, it is

not applicable to the assumed case of positive semi-definite $U(\cdot, 0)$, because, the value function is positive definite. The following lemma pursues another idea to this end.

Lemma 3. *If $h(\cdot)$ is an admissible control policy within Ω , then $V_h(\cdot) \in \mathcal{C}(\Omega)$.*

Proof: The proof is done by contradiction. Assume that $V_h(\cdot)$ is discontinuous at some $y_0 \in \Omega$. Then

$$\exists \epsilon > 0, \forall \delta > 0, \exists x_0 \in \Omega : \quad (24)$$

$$|V_h(x_0) - V_h(y_0)| > \epsilon \text{ while } \|x_0 - y_0\| < \delta,$$

where $\|\cdot\|$ denotes a vector norm and ‘ \cdot ’ denoted ‘such that’. The idea is showing that (24) is not possible. To this end, considering recursive relation (21) initiated with $V_h^0(\cdot) = 0$, from (3) and (22) one has

$$V_h(x_0) = V_h^j(x_0) + V_h(x_j^h). \quad (25)$$

Therefore,

$$V_h(x_0) - V_h(y_0) = V_h^j(x_0) + V_h(x_j^h) - V_h^j(y_0) - V_h(y_j^h), \quad (26)$$

which leads to

$$|V_h(x_0) - V_h(y_0)| \leq |V_h^j(x_0) - V_h^j(y_0)| + V_h(x_j^h) + V_h(y_j^h), \quad (27)$$

by triangle inequality of absolute values. Inequality (27) is the key to the solution, as it will be shown that the right hand side of the inequality can be made arbitrarily small if x_0 is close enough to y_0 and j is large enough. By $V_h(x) \leq W(x), \forall x$, for some $W(\cdot) \in \mathcal{C}(\Omega), W(0) = 0$, per the admissibility of $h(\cdot)$, one has

$$|V_h(x_0) - V_h(y_0)| \leq |V_h^j(x_0) - V_h^j(y_0)| + W(x_j^h) + W(y_j^h). \quad (28)$$

By admissibility of $h(\cdot)$ one has $y_j^h \rightarrow 0$ as $j \rightarrow \infty$. Hence, by $W(\cdot) \in \mathcal{C}(\Omega)$ and $W(0) = 0$, one has

$$\forall y_0 \in \Omega, \forall \epsilon > 0, \exists j_1 = j_1(y_0, \epsilon) : \quad (29)$$

$$j \geq j_1 \Rightarrow 0 \leq W(y_j^h) < \epsilon/4.$$

Moreover, by $W(\cdot) \in \mathcal{C}(\Omega)$

$$\forall y_j^h \in \Omega, \forall \epsilon > 0, \exists \delta_1 = \delta_1(y_j^h, \epsilon) : \quad (30)$$

$$\|x_j^h - y_j^h\| < \delta_1 \Rightarrow |W(x_j^h) - W(y_j^h)| < \epsilon/4.$$

Hence,

$$\|x_j^h - y_j^h\| < \delta_1 \Rightarrow W(x_j^h) < \epsilon/4 + W(y_j^h). \quad (31)$$

On the other hand, due to the Lipschitz continuity of the closed loop system, the state trajectory at each finite time, for example j , continuously depends on the initial conditions, [25]. In other words, x_j^h changes continuously as x_0 changes, hence,

$$\forall y_0 \in \Omega, \forall \delta_1 > 0, \exists \delta_2 = \delta_2(y_0, \delta_1, j) : \quad (32)$$

$$\|x_0 - y_0\| < \delta_2 \Rightarrow \|x_j^h - y_j^h\| < \delta_1.$$

Note that $V_h^j(\cdot) \in \mathcal{C}(\Omega)$, for $j < \infty$, as mentioned before this lemma. Hence,

$$\forall y_0 \in \Omega, \forall \epsilon > 0, \exists \delta_3 = \delta_3(y_0, \epsilon, j) : \quad (33)$$

$$\|x_0 - y_0\| < \delta_3 \Rightarrow |V_h^j(x_0) - V_h^j(y_0)| < \epsilon/4.$$

Now we have enough inequalities to contradict (24). For any point of discontinuity y_0 and ϵ whose existence is guaranteed by (24), find $j_1 = j_1(y_0, \epsilon)$ which leads to

$$W(y_{j_1}^h) < \epsilon/4, \quad (34)$$

per (29). Then, select $\delta_1 = \delta_1(y_{j_1}^h, \epsilon)$. Per (31) and (34) one has

$$\|x_{j_1}^h - y_{j_1}^h\| < \delta_1 \Rightarrow W(x_{j_1}^h) < \epsilon/4 + W(y_{j_1}^h) < \epsilon/2. \quad (35)$$

Select $\delta_2 = \delta_2(y_0, \delta_1, j_1)$ to have

$$\|x_0 - y_0\| < \delta_2 \Rightarrow \|x_{j_1}^h - y_{j_1}^h\| < \delta_1, \quad (36)$$

per (32). Finally, set $\delta_3 = \delta_3(y_0, \epsilon, j_1)$ to have

$$\|x_0 - y_0\| < \delta_3 \Rightarrow |V_h^{j_1}(x_{j_1}^h) - V_h^{j_1}(y_{j_1}^h)| < \epsilon/4, \quad (37)$$

per (33). Select $\delta = \min(\delta_2, \delta_3)$. Using (34), (35), (36), and (37) one has

$$\|x_0 - y_0\| < \delta \Rightarrow |V_h(x_0) - V_h(y_0)| \leq$$

$$|V_h^{j_1}(x_0) - V_h^{j_1}(y_0)| + W(x_{j_1}^h) + W(y_{j_1}^h) < \epsilon, \quad (38)$$

which contradicts (24). Therefore, $V_h(\cdot) \in \mathcal{C}(\Omega)$. \square

Even though we managed to skip the proof of uniform convergence for deriving the desired continuity result, the uniformness of the convergence of (21), if established, will still be useful. The reason is, the uniform convergence can be used in finding the value function of an admissible control through guaranteeing that there exists a large enough iteration index such that one has

$$|V_h^j(x) - V_h^{j+1}(x)| \leq \delta, \forall x \in \Omega, \quad (39)$$

for any selected *constant* tolerance $\delta > 0$. This condition can be used for terminating the iterations. The next lemma proves the desired uniform convergence.

Lemma 4. *If $h(\cdot)$ is an admissible control policy within compact set Ω , then selecting any $V_h^0(\cdot) \in \mathcal{C}(\Omega)$ which satisfies $0 \leq V_h^0(x) \leq U(x, 0), \forall x \in \Omega$, the iterations given by (21) converges uniformly in Ω .*

Proof: Using Dini’s uniform convergence theorem (Ref. [26], Theorem 7.13), the pointwise monotonicity of $\{V_h^j(x)\}_{j=0}^\infty$ (Lemma 2), the continuity of the elements of the foregoing sequence, the continuity of the limit function $V_h(\cdot)$ (Lemma 3), and the compactness of Ω , lead to the uniform convergence of the iterations. \square

Now that the concern about continuity of $V_h(\cdot)$ is resolved, the next step is a lemma which proves that the *argmin* operator will not cause discontinuity in the functions subject to investigation.

Lemma 5. *Let $W(x, u) := U(x, u) + V^i(f(x, u))$ and $h(x) = \text{argmin}_{u \in \mathbb{R}^m} W(x, u)$. If functions $f(\cdot, \cdot), U(\cdot, \cdot)$, and $V^i(\cdot)$ are continuous within Ω , then so is $W(\cdot, h(\cdot))$.*

Proof: The proof is done by showing that the directional limit of $W(\cdot, h(\cdot))$ at any selected point is equal to its evaluation at the point, and hence, it is continuous at that point (motivated by [31]).

Let \bar{x} be an arbitrary point in Ω . Set

$$\bar{u} := h(\bar{x}). \quad (40)$$

Select an open set $\alpha \subset \mathbb{R}^n$ such that \bar{x} belongs to the boundary of α and limit

$$\hat{u} := \lim_{x \rightarrow \bar{x}, x \in \alpha} h(x), \quad (41)$$

exists. If $\bar{u} = \hat{u}$, for every such α , then $h(\cdot) \in \mathcal{C}(\bar{x})$. In

this case the continuity of $W(\cdot, h(\cdot))$ at \bar{x} follows from the continuity of its forming functions, [26].

Now assume $\bar{u} \neq \hat{u}$, for some α denoted with α_0 . From $W(\cdot, \hat{u}) \in \mathcal{C}(\Omega)$ for the given \hat{u} , one has

$$W(\bar{x}, \hat{u}) = \lim_{x \rightarrow \bar{x}, x \in \alpha_0} W(x, \hat{u}), \quad (42)$$

If it can be shown that, for every selected α_0 , one has

$$W(\bar{x}, \bar{u}) = W(\bar{x}, \hat{u}), \quad (43)$$

then the continuity of $W(\cdot, h(\cdot))$ at \bar{x} follows, because from (42) and (43) one has

$$W(\bar{x}, \bar{u}) = \lim_{x \rightarrow \bar{x}} W(x, \hat{u}), \quad (44)$$

and (44) leads to the continuity by definition, [26].

The proof that (43) holds is done by contradiction. Assume that for some \bar{x} and some α_0 one has

$$W(\bar{x}, \bar{u}) > W(\bar{x}, \hat{u}). \quad (45)$$

Inequality (45) leads to $h(\bar{x}) \neq \bar{u}$. But, this is against (40), hence, (45) cannot hold. Now, assume

$$W(\bar{x}, \bar{u}) < W(\bar{x}, \hat{u}), \quad (46)$$

hence there exists some $\epsilon_1 > 0$ such that

$$W(\bar{x}, \bar{u}) + \epsilon_1 = W(\bar{x}, \hat{u}), \quad (47)$$

then, due to the continuity of both sides of (47) at \bar{x} for the fixed \bar{u} and \hat{u} , there exists an open set γ containing \bar{x} , see Fig. 1, and some $\epsilon_2 > 0$, such that

$$W(x, \bar{u}) + \epsilon_2 < W(x, \hat{u}), \forall x \in \gamma. \quad (48)$$

Given $W(x, h(x)) \leq W(x, \bar{u})$, inequality (48) implies that at points which are *close enough* to \bar{x} , function $W(x, h(x))$ is away from $W(x, \hat{u})$ at least by a margin of ϵ_2 . But, this contradicts Eq. (41) which, implies that $h(x)$ can be made arbitrarily close to \hat{u} as x gets close to \bar{x} within α_0 . The reason is, the latter, given the continuity of $W(x, u)$ versus both x and u , leads to the conclusion that function $W(x, h(x))$ can be made arbitrarily close to $W(x, \hat{u})$ if x approaches \bar{x} from a certain direction. Note that sets γ and α_0 are not disjoint, as \bar{x} is *within* γ and on the *boundary* of α_0 , as shown in Fig. 1. Hence, inequality (46) also cannot hold. Therefore, (43) holds and hence, $W(\cdot, h(\cdot)) \in \mathcal{C}(\bar{x})$. Finally, the continuity of the function subject to investigation at any arbitrary $\bar{x} \in \Omega$, leads to the continuity of the function in Ω . \square

Now, we have all the required tools to make the following desired conclusion.

Theorem 1. *The value functions at each iteration of stabilizing value iteration are continuous functions, i.e., $V^i(\cdot) \in \mathcal{C}(\Omega)$, $\forall i \in \mathbb{N}$.*

Proof: The theorem can be proved by induction. Lemma 3 proves $V^0(\cdot) \in \mathcal{C}(\Omega)$. Assume that $V^i(\cdot) \in \mathcal{C}(\Omega)$. From Lemma 5 it follows that $V^{i+1}(\cdot) \in \mathcal{C}(\Omega)$, because, $W(\cdot, h(\cdot)) = V^{i+1}(\cdot)$ where $W(\cdot, h(\cdot))$ is defined in Lemma 5. \square

The proof of continuity is of interest for two reasons. 1) to guarantee suitable approximation capability, especially in

generalization, i.e., approximating the function at the sample states which were not used during the training stage, [28], [29], and 2) for using the value functions as Lyapunov functions, for proof of stability, as given next.

Theorem 2. *Let the compact domain \mathcal{B}_r^i for any $r \in \mathbb{R}_+$ be defined as $\mathcal{B}_r^i := \{x \in \mathbb{R}^n : V^i(x) \leq r\}$ and let $\bar{r}^i > 0$ be (the largest r) such that $\mathcal{B}_{\bar{r}^i}^i \subset \Omega$. Then, for every given $i \in \mathbb{N}$, control policy $h^i(\cdot)$ resulting from stabilizing value iteration asymptotically stabilizes the system about the origin and $\mathcal{B}_{\bar{r}^i}^i$ will be an estimation of the region of attraction for the system.*

Proof: The proof is done by showing that $V^i(\cdot)$ is a Lyapunov function for $h^i(\cdot)$, for each given i . Denoting the value function of the initial stabilizing control with $V^0(\cdot)$, it is continuous (Lemma 3) and positive definite, by positive semi-definiteness of $U(\cdot, \cdot)$ and Assumption 2. Note that, there is no $x \neq 0$ with the value function of zero under any control policy. If $V^i(\cdot)$ for some i is positive definite, it directly follows from (10) that $V^{i+1}(\cdot)$ will also be positive definite, because, if $U(x, 0) = 0$ for some $x \neq 0$, then $f(x, 0) \neq x$ by Assumption 2. Hence, by induction, $V^{i+1}(\cdot)$ is positive definite for every $i \in \mathbb{N}$. Also, by Theorem 1 it is a continuous function in Ω .

By (10)

$$V^i(f(x, h^i(x))) - V^{i+1}(x) = -U(x, h^i(x)), \forall x \in \Omega. \quad (49)$$

On the other hand, by Lemma 1, inequality (20) holds for all x . Therefore, replacing $V^{i+1}(x)$ in (49) with $V^i(x)$ leads to

$$V^i(f(x, h^i(x))) - V^i(x) \leq -U(x, h^i(x)), \forall x \in \Omega. \quad (50)$$

Let $S := \{x \in \mathbb{R}^n : U(x, 0) = 0\}$. The right hand side of (50) can be zero only if $x \in S$. Since, by Assumption 2, no non-zero state trajectory can stay in S , the asymptotic stability of $h^i(\cdot)$ follows from negative semi-definiteness of the difference between the value functions in (50), using LaSalle's invariance theorem, [25].

Set $\mathcal{B}_{\bar{r}^i}^i$ is an EROA [25] for the closed loop system, because, $V^i(x_{k+1}) \leq V^i(x_k)$ by (50), hence, $x_k \in \mathcal{B}_{\bar{r}^i}^i$ leads to $x_{k+1} \in \mathcal{B}_{\bar{r}^i}^i, \forall k \in \mathbb{N}$. Finally, since $\mathcal{B}_{\bar{r}^i}^i$ is contained in Ω , it is bounded. Also, the set is closed, because, it is the *inverse image* of a closed set, namely $[0, \bar{r}^i]$ under a continuous function (Theorem 1), [26]. Hence, $\mathcal{B}_{\bar{r}^i}^i$ is compact. The origin is an *interior* point of the EROA, because $V^i(0) = 0$, $\bar{r}^i > 0$, and $V^i(\cdot) \in \mathcal{C}(\Omega)$. \square

Theorem 2 proves that each single $h^i(\cdot)$ if *constantly* applied on the system, will have the states converge to the origin. However, in online learning, the control will be subject to adaptation. In other words, if $h^i(\cdot)$ is applied at the current time, control policy $h^{i+1}(\cdot)$ will be applied at the next time-step. It is important to note that even though Theorem 2 proves the asymptotic stability of the *autonomous* system $x_{k+1} = F(x_k) := f(x_k, h^i(x_k))$ for every fixed i , it does not guarantee the asymptotic stability of the *non-autonomous* system $x_{k+1} = F(x_k, k) := f(x_k, h^k(x_k))$. Therefore, it is required to have a separate stability analysis to show that the trajectory formed under the *adapting/evolving* control policy also will converge to zero. An idea for doing that is finding a *single* function, possibly $V^0(\cdot)$, to be a Lyapunov function for *all* the control policies. The proof of the following theorem,

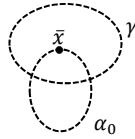


Fig. 1. Schematic of point \bar{x} and open sets γ and α_0 .

however, uses another approach.

Theorem 3. *If the system is operated using control policy $h^k(\cdot)$ at time k , that is, the control subject to adaptation in the stabilizing value iteration, then, the origin will be asymptotically stable and every trajectory contained in Ω will converge to the origin.*

Proof: Eq. (10) and the monotonicity feature established in Lemma 1 lead to

$$V^1(x_0) = U(x_0, h^0(x_0)) + V^0(f(x_0, h^0(x_0))) \leq V^0(x_0), \forall x_0 \in \Omega, \quad (51)$$

and similarly

$$V^2(x) = U(x, h^1(x)) + V^1(f(x, h^1(x))) \leq V^1(x) \leq V^0(x), \forall x \in \Omega. \quad (52)$$

Let $x_k^* := f(x_{k-1}^*, h^{k-1}(x_{k-1}^*))$ for $k \in \mathbb{N} - \{0\}$, and $x_0^* := x_0$. Evaluating (52) at x_1^* and replacing the $V^0(x_1^*)$ in the left hand side of the inequality in (51) with the left hand side of (52), which is smaller per (52), one has

$$U(x_0^*, h^0(x_0^*)) + U(x_1^*, h^1(x_1^*)) + V^1(x_2^*) \leq V^0(x_0^*), \forall x_0^* \in \Omega. \quad (53)$$

Repeating this process by replacing $V^1(x_2^*)$ in (53) using

$$V^3(x_2^*) = U(x_2^*, h^2(x_2^*)) + V^2(x_3^*) \leq V^2(x_2^*) \leq V^1(x_2^*), \forall x_2^* \in \Omega. \quad (54)$$

leads to

$$U(x_0^*, h^0(x_0^*)) + U(x_1^*, h^1(x_1^*)) + U(x_2^*, h^2(x_2^*)) + V^2(x_3^*) \leq V^0(x_0^*), \forall x_0^* \in \Omega. \quad (55)$$

Similarly by repeating this process one has

$$\sum_{k=0}^{i-1} U(x_k^*, h^k(x_k^*)) + V^{i-1}(x_i^*) \leq V^0(x_0^*), \quad \forall x_0^* \in \Omega, \forall i \in \mathbb{N} - \{0\}. \quad (56)$$

Since $V^{i-1}(x) \geq 0, \forall x$, the foregoing equations leads

$$\sum_{k=0}^{i-1} U(x_k^*, h^k(x_k^*)) \leq V^0(x_0^*), \forall x_0^* \in \Omega, \forall i \in \mathbb{N} - \{0\}, \quad (57)$$

that is, the sequence of partial sums of the left hand side is upper bounded and because of being non-decreasing, it converges, as $i \rightarrow \infty$, [26]. Therefore, $U(x_i^*, h^i(x_i^*)) \rightarrow 0$ as $i \rightarrow \infty$. Considering Assumption 2, this leads to $x_i^* \rightarrow 0$, as long as the entire state trajectory is contained in Ω . \square

It can be seen that Theorem 3 does not provide an EROA. Therefore, the training domain Ω needs to be selected large enough to guarantee that states, on their way of traveling toward the origin (with not a necessarily straight path) do not exit Ω . However, once the convergence of the value iteration is established, some analytical results regarding the desired EROA will be presented (Theorem 5).

Besides stability, which is addressed, the convergence of the VI using the stabilizing initial guess also needs to be analyzed. The reason is, none of the cited existing convergence proofs is applicable, as they either require $V^0(x) = 0, \forall x$, [10], or $0 \leq V^0(x) \leq U(x, 0), \forall x$, [20]. For example, $0 \leq V^0(x) \leq U(x, 0), \forall x$, does not hold here because $U(x, \cdot)$,

which is greater than or equal to $U(x, 0)$, is only one of the terms existing in the summation over infinite number of non-negative terms in the definition of $V^0(x)$ as a value function of state x . As for the convergence result in [18] whose less restrictive version was presented in [32], it requires $V^*(f(x, u)) \leq \beta U(x, u)$ to hold uniformly. This condition is restrictive and not applicable to our analysis, as detailed before Lemma 3.

The main idea for the convergence proof in here is adapted from [20], in which, an analogy was established between the iterations of the VI and the *horizon length* of a finite-horizon optimal control problem with a *fixed final time*. Let the cost function for the finite-horizon problem be given by

$$J^N = \psi(x_N) + \sum_{k=0}^{N-1} U(x_k, u_k), \quad (58)$$

where $\psi : \mathbb{R}^n \rightarrow \mathbb{R}_+$ is a continuous and positive semi-definite function representing the *terminal cost*. The finite-horizon problem is defined as minimizing J^N subject to the dynamics given by (1). Once the final time is fixed, the value function and the control policy become time-dependent [30], [13], i.e., they may be denoted with $V^*(\cdot, \cdot)$ and $h^*(\cdot, \cdot)$, respectively, where the second argument is the number of remaining time steps, or *time-to-go*. Let the optimal finite-horizon value function given state x_0 and time-to-go $\tau \in \mathbb{M} := \{0, 1, \dots, N\}$ be denoted by $V^* : \mathbb{R}^n \times \mathbb{M} \rightarrow \mathbb{R}_+$, where

$$V^*(x_0, \tau) = \psi(x_\tau) + \sum_{k=0}^{\tau-1} U(x_k^{*,\tau}, h^*(x_k^{*,\tau}, \tau - k)), \quad (59)$$

$x_k^{*,\tau} := f(x_{k-1}^{*,\tau}, h^*(x_{k-1}^{*,\tau}, \tau - (k-1)))$, $\forall k$ such that $1 \leq k \leq \tau$, and $x_0^{*,\tau} := x_0, \forall \tau$. In other word, the summation is evaluated along the trajectory generated by applying the time varying control policy $h^*(\cdot, \tau - k)$ at time k . Clearly

$$V^*(x, 0) = \psi(x), \forall x, \quad (60)$$

and by the Bellman equation for fixed-final-time problems [30], [13]

$$V^*(x, \tau + 1) = \min_{u \in \mathbb{R}^m} (U(x, u) + V^*(f(x, u), \tau)), \quad \forall x, \forall \tau \in \mathbb{M} - \{N\}, \quad (61)$$

and

$$h^*(x, \tau + 1) = \operatorname{argmin}_{u \in \mathbb{R}^m} (U(x, u) + V^*(f(x, u), \tau)), \quad \forall x, \forall \tau \in \mathbb{M} - \{N\}. \quad (62)$$

If $\psi(\cdot)$ in the finite-horizon problem is selected equal to initial guess $V^0(\cdot)$ in the VI, then, comparing Eq. (61) with (11) it directly follows that

$$V^*(x, i) = V^i(x), \forall x, \forall i \in \mathbb{M}. \quad (63)$$

In other words, the *immature* value function at the i th iteration of VI is identical to the *optimal* value function of the fixed-final-time problem of minimizing (58) with the final time of i , when $\psi(\cdot) = V^0(\cdot)$. Similarly comparing (62) with (9) it can be seen that

$$h^*(x, i + 1) = h^i(x), \forall x, \forall i \in \mathbb{M}. \quad (64)$$

Using this idea, it is proved in [20] that if $V^0(\cdot)$ is smooth

and $0 \leq V^0(x) \leq U(x, 0), \forall x$, then VI converges to $V^*(\cdot)$. The following theorem generalizes the convergence proof to cover the case of $V^0(\cdot)$ being a value function.

Before proceeding to the theorem, however, it is noteworthy that considering Eq. (63), the stability results given by Theorem 2 resembles a method of stability proof in the model predictive control (MPC) literature [33], in which, a control Lyapunov function is utilized as the terminal cost in the respective finite-horizon problems. Using this idea, it is shown that the closed loop system, under each control calculated for the receding horizons, remains stable.

Theorem 4. *The stabilizing value iteration converges to the optimal solution of the infinite-horizon problem within the selected compact domain.*

Proof: Considering the analogy between the iteration of VI and the horizon of a finite-horizon problem given by (63), it can be seen that each $V^i(x_0)$ represents the cost-to-go of applying control sequence $\{h^*(x_0, i), h^*(x_1, i-1), \dots, h^*(x_{i-1}, 1)\}$ for the first i steps, which are the optimal control sequence with respect to cost function (58), when $\psi(\cdot) = V^0(\cdot)$ and $N = i$, and applying the stabilizing control sequence $\{h^{-1}(x_i), h^{-1}(x_{i+1}), \dots\}$ for the rest of the (infinite) horizon. The reason for this conclusion is the fact that $V^0(x_i)$, which is used as the terminal cost in the fixed-final-time problem itself represents the cost-to-go of applying admissible control policy $h^{-1}(\cdot)$ for infinite number of times, starting from x_i , per Eq. (3).

On the other hand, the non-increasing (cf. Lemma 1) and non-negative (cf. proof of Theorem 2) nature of value functions under VI, and hence, of the finite-horizon value function $V^*(\cdot, i)$ lead to the convergence of the sequence of value functions to a finite limit function, denoted with $V^\infty(\cdot) = V^*(\cdot, \infty)$. Because, every non-increasing and lower bounded sequence converges, [26]. Therefore, one has

$$\lim_{i \rightarrow \infty} x_i \rightarrow 0, \quad (65)$$

using control sequence $\{h^*(x_0, i), h^*(x_1, i-1), \dots, h^*(x_{i-1}, 1)\}$. Otherwise, $V^\infty(\cdot)$ becomes unbounded. This can also be concluded by noting that the tail of a convergent series can be made arbitrarily small (cf. p. 59 [26]). Note that per Assumption 2 the state trajectory cannot hide in the invariant set of $f(\cdot, 0)$ with zero utility function, to lead to a finite cost-to-go without convergence to the origin.

Due to the continuity and positive semi-definiteness of $V^0(\cdot)$, one has $V^0(\cdot) \rightarrow 0$ as $x \rightarrow 0$. Therefore, by Eq. (65) one has

$$\lim_{i \rightarrow \infty} V^0(x_i) \rightarrow 0, \quad (66)$$

in calculation of the cost-to-go $V^*(\cdot, i)$. Comparing finite-horizon cost function (58) with infinite-horizon cost function (2) and considering (66), one has

$$V^*(x) = V^\infty(x), \forall x \in \Omega. \quad (67)$$

Otherwise, the smaller value among $V^*(x)$ and $V^\infty(x)$ will be both the optimal value function (evaluated at x) for the infinite-horizon problem and the greatest lower bound of the sequence of value function of the fixed-final-time problems resulting from $N = 0, 1, 2, \dots$ \square

Comparing the results given by Theorems 2 and 4 with the existing literature, the closest one is [24], in which a VI algorithm, called θ -ADP, was introduced. The point that θ -ADP requires to be initiated from a control Lyapunov function (CLF) of the respective system in order for the control under iterations to be stabilizing corresponds to the required initial admissible guess for VI in this study. The reason is, if the CLF is known, an asymptotically stable control can be derived directly from the function, e.g., using Sontag's formula, [34]. An asymptotically stable control law, however, is not required to lead to a finite cost-to-go. Hence, at the first glance, the condition in θ -ADP seems to be less restrictive compared to the condition of using admissible initial control in this study. However, once the second requirement of θ -ADP, that is the existence and utilization of a scale factor θ using which the CLF function evaluated along the state trajectory decays faster than a value function, is taken into account, the control resulting from the scaled CLF will lead to a finite cost-to-go. Therefore, the results look similar in regards to the initial guess. However, the simplicity of the proofs especially for the convergence proof, including the intermediate steps for firm conclusions (e.g., continuity analysis), admitting a positive semi-definite running cost as opposed to the positive definite one in that work, and establishing an EROA are the main differences of the mentioned theorems compared with [24].

Besides addressing the convergence concern in VI, the foregoing theorem provides an idea for establishing an EROA for the system operated using evolving control policies, as presented next.

Theorem 5. *Let $\mathcal{B}_r^i := \{x \in \mathbb{R}^n : V^i(x) \leq r\}$ and $\mathcal{B}_r^* := \{x \in \mathbb{R}^n : V^*(x) \leq r\}$ for any $r \in \mathbb{R}_+$. Also, let the system be operated using control policy $h^k(\cdot)$ at time k , that is, the control subject to adaptation in the stabilizing value iterations. If $\mathcal{B}_r^* \subset \Omega$ for an $r > 0$ then \mathcal{B}_r^0 is an estimation of the region of attraction of the closed loop system.*

Proof: As the first step we show that for any given r one has

$$x_k \in \mathcal{B}_r^k \Rightarrow x_{k+1} = f(x_k, h^k(x_k)) \in \mathcal{B}_r^{k+1}, \forall k \in \mathbb{N}, \forall r \in \mathbb{R}_+. \quad (68)$$

By inequality (50) one has $V^k(x_{k+1}) \leq V^k(x_k)$. Therefore,

$$x_k \in \mathcal{B}_r^k \Rightarrow x_{k+1} = f(x_k, h^k(x_k)) \in \mathcal{B}_r^k, \forall k \in \mathbb{N}, \forall r \in \mathbb{R}_+. \quad (69)$$

By definition of \mathcal{B}_r^k if $V^{k+1}(x) \leq V^k(x), \forall x$, which follows from Lemma 1, then $\mathcal{B}_r^k \subset \mathcal{B}_r^{k+1}$. Therefore,

$$x_{k+1} \in \mathcal{B}_r^k \Rightarrow x_{k+1} \in \mathcal{B}_r^{k+1}, \forall k \in \mathbb{N}, \forall r \in \mathbb{R}_+. \quad (70)$$

Finally (69) and (70) lead to (68). Now that (68) is proved, one may use mathematical induction to see

$$x_0 \in \mathcal{B}_r^0 \Rightarrow x_k \in \mathcal{B}_r^k, \forall k \in \mathbb{N}, \forall r \in \mathbb{R}_+. \quad (71)$$

The next step is noticing that $V^*(x) \leq V^i(x), \forall x$, which follows from the monotonicity of $\{V^i(\cdot)\}_{i=0}^\infty$ and its convergence to $V^*(\cdot)$, per Theorem 4. The foregoing inequality leads to $\mathcal{B}_r^k \subset \mathcal{B}_r^*, \forall k$, by definition of \mathcal{B}_r^k and \mathcal{B}_r^* . Therefore, (71) leads to

$$x_0 \in \mathcal{B}_r^0 \Rightarrow x_k \in \mathcal{B}_r^*, \forall k \in \mathbb{N}, \forall r \in \mathbb{R}_+. \quad (72)$$

The result given by (72) proves the theorem, because, if r is such that $\mathcal{B}_r^* \subset \Omega$ then any trajectory initiated within \mathcal{B}_r^0 will remain inside Ω , and hence, by Theorem 3 will converge to the origin. \square

V. ANALYSIS OF STABILIZING APPROXIMATE VALUE ITERATION

The problem with the exact VI is the issue that *exact* reconstruction of the right hand side of Eq. (11) is not generally possible except for very simple problems. In general, parametric function approximators are used for this purpose, which hence, give rise to function approximation errors. When the approximation errors are considered, Eq. (11) reads

$$\hat{V}^{i+1}(x) = \min_{u \in \mathbb{R}^m} \left(U(x, u) + \hat{V}^i(f(x, u)) \right) + \epsilon^i(x), \forall x \in \Omega, \quad (73)$$

where the *approximate value function* at the i th iteration is denoted with $\hat{V}^i(\cdot)$ and the approximation error at this iteration is denoted with $\epsilon^i(\cdot)$. Note that the value function in the right hand side of Eq. (73) is also an approximate quantity, generated from the previous iteration. When $\epsilon^i(\cdot) \neq 0$, the convergence of the approximate VI (AVI) does not follow from Theorem 4. This convergence/boundedness is investigated in this section.

Before proceeding to the convergence/boundedness analysis it is worth mentioning that one typically trains a control approximator (actor) to approximate the solution to the minimization problem given by (5) based on the value function resulting from VI. The control approximator, will hence, lead to *another* approximation error term in the process, regardless of whether the value function reconstruction is exact or approximate. However, the effect of the actor's approximation error can be removed from both the convergence analysis of the AVI and the stability analysis of the system during value iterations, as the control will be directly calculated from the minimization of the right hand side of Eq. (73) in *online* and *adaptive* optimal control and applied on the system. In other words, even though the actor will be updated simultaneously along with the critic in online learning, the critic training and the system's operation are independent of the actor's approximation accuracy. Once the learning is concluded (and if it is concluded), the operation of the system could be based on the control resulting from the trained actor, hence, the actor's approximation error can affect the stability of the system at that stage. The stability analysis *after* conclusion of AVI is beyond the scope of this study and deserves to be investigated separately, as the focus in this work is analyzing the convergence/boundedness and stability *during* the online learning process through AVI.

Considering the above comment and denoting the minimizer of the right hand side of Eq. (73) by $\hat{h}^i(\cdot)$, one has

$$\hat{h}^i(x) = \operatorname{argmin}_{u \in \mathbb{R}^m} \left(U(x, u) + \hat{V}^i(f(x, u)) \right), \forall x \in \Omega, \quad (74)$$

therefore, Eq. (73) can be written as

$$\hat{V}^{i+1}(x) = U(x, \hat{h}^i(x)) + \hat{V}^i(f(x, \hat{h}^i(x))) + \epsilon^i(x), \forall x \in \Omega. \quad (75)$$

Assuming an upper bound for the approximation error $\epsilon^i(x)$ the results given by Theorem 6 can be obtained, in terms of boundedness of sequence $\{\hat{V}^i(x)\}_{i=0}^\infty$ resulting from the AVI and its relation versus the optimal value function. This boundedness will later be used for stability analysis.

For our AVI analyses, i.e., for the rest of this study, it is assumed that state penalizing function $Q(\cdot)$ in $U(x, u) = Q(x) + u^T R u$ only vanishes at the origin. In other words, instead of the positive semi-definiteness of $U(\cdot, 0)$, it is assumed that $U(\cdot, 0)$ is positive definite hereafter. The reason for this modification is the point that $U(x, 0)$ is going to be used to put an upper bound on $|\epsilon^i(x)|$, which makes sense only if it does not vanish at any non-zero x .

Theorem 6. *Let $|\epsilon^i(x)| \leq cU(x, 0), \forall x \in \Omega, \forall i \in \mathbb{N}$, for some $c \in [0, 1)$. If the approximate value iteration is initiated using some $\hat{V}^0(x)$ which satisfies $\underline{V}^0(x) \leq \hat{V}^0(x) \leq \bar{V}^0(x), \forall x \in \Omega$ where $\underline{V}^0(x)$ and $\bar{V}^0(x)$ are, respectively, the initial guesses for the exact value iterations corresponding to cost functions*

$$\underline{J} = \sum_{k=0}^{\infty} \left(U(x_k, u_k) - cU(x_k, 0) \right), \quad (76)$$

and

$$\bar{J} = \sum_{k=0}^{\infty} \left(U(x_k, u_k) + cU(x_k, 0) \right), \quad (77)$$

subject to dynamics (1), then, the result of the approximate value iteration at the i th iteration is bounded from below by the result of the exact value iteration corresponding to cost function (76) and from above by the result of the exact value iteration corresponding to cost function (77).

Proof: Let $\{\bar{V}^i(x)\}_{i=0}^\infty$ and $\{\underline{V}^i(x)\}_{i=0}^\infty$ where $\bar{V}^i : \mathbb{R}^n \rightarrow \mathbb{R}_+$ and $\underline{V}^i : \mathbb{R}^n \rightarrow \mathbb{R}_+$, be defined as sequences of functions initiated from some $\bar{V}^0(\cdot)$ and $\underline{V}^0(\cdot)$ and generated by

$$\underline{V}^{i+1}(x) = \min_{u \in \mathbb{R}^m} \left(U(x, u) - cU(x, 0) + \underline{V}^i(f(x, u)) \right), \forall x \in \Omega, \quad (78)$$

$$\bar{V}^{i+1}(x) = \min_{u \in \mathbb{R}^m} \left(U(x, u) + cU(x, 0) + \bar{V}^i(f(x, u)) \right), \forall x \in \Omega. \quad (79)$$

Considering recursive relations (79) and (78) it is seen that $\bar{V}^i(\cdot)$ and $\underline{V}^i(\cdot)$ are, respectively, the value functions at the i th iteration of *exact* VI for cost functions (76) and (77). Considering this point, the lemma can be proved using mathematical induction. Initially $\hat{V}^0(x) \leq \bar{V}^0(x), \forall x \in \Omega$ by assumption. Let $\hat{V}^i(x) \leq \bar{V}^i(x), \forall x \in \Omega$ hold for some i . Comparing Eq. (79) with Eq. (73) it follows that $\hat{V}^{i+1}(x) \leq \bar{V}^{i+1}(x)$, since $\epsilon^i(x) \leq cU(x, 0)$ and $\hat{V}^i(x) \leq \bar{V}^i(x), \forall x$. Therefore, one has $\hat{V}^i(x) \leq \bar{V}^i(x), \forall i \in \mathbb{N}, \forall x$. The proof of $\underline{V}^i(x) \leq \hat{V}^i(x), \forall i \in \mathbb{N}$ is similar by induction, through comparing Eq. (78) with Eq. (73) and noting that $-cU(x, 0) \leq \epsilon^i(x), \forall x, \forall i$. \square

The result given by the foregoing theorem resembles the idea of Relaxed Dynamic Programming presented in [18]. However, the idea, the proof, and the applications of the boundedness result, presented in the rest of this study, are

different.

The exact VIs given by (78) and (79) converge, based on Theorem 4, when initiated using value functions (defined based on the respective cost function) of some admissible controls. Therefore, considering Theorem 6, the boundedness of the AVI results for all iterations follows, assuming the boundedness of the approximation errors by $cU(x, 0)$ for some $c \in [0, 1)$.

The actual convergence as well as stability of the system operated under AVI are much more challenging compared to the respective analyses in exact VI, since the presence of the approximation error cancels the monotonicity feature presented in Lemma 1. Note that the monotonicity was the backbone of both the stability and the convergence results given in Theorems 2, 3, and 4. As long as the boundedness of the functions during AVI is guaranteed in a neighborhood of the optimal value function (Theorem 6) where the neighborhood shrinks if the approximation error decreases, the actual convergence of the iteration may not be of a critical importance in implementing the AVI. But, the stability of the system operated under the AVI is definitely critical. The following lemma develops a ‘semi-monotonicity’ of the stabilizing AVI to be used later for deriving some stability results. Note that, following Definition 3, *stabilizing AVI* is defined as the AVI which is initiated using the *approximate* value function of an admissible control policy, with an approximation error denoted with $\epsilon^{-1}(\cdot)$. In other words instead of the exact value function $V^0(\cdot)$ given by (13) one initiates the iterations using the approximate value function $\hat{V}^0(\cdot)$ which satisfies

$$\hat{V}^0(x) = U(x, h^{-1}(x)) + \hat{V}^0(f(x, h^{-1}(x))) + \epsilon^{-1}(x), \forall x \in \Omega. \quad (80)$$

Lemma 6. Let $\check{V}^i(x_0) := \sum_{k=0}^i U(\hat{x}_k^{*,i}, 0)$, $\forall x_0 \in \Omega$, $\forall i \in \mathbb{N}$, where $\hat{x}_k^{*,i} := f(\hat{x}_{k-1}^{*,i}, \hat{h}^{i-k}(\hat{x}_{k-1}^{*,i}))$ and $\hat{x}_0^{*,i} := x_0$, $\forall i, \forall k \in \mathbb{N} - \{0\}$. If the stabilizing approximate value iteration scheme is conducted using a function approximator that satisfies $|\epsilon^i(x)| \leq cU(x, 0)$, $\forall i \in \mathbb{N} \cup \{-1\}$, $\forall x$, for some $c \in [0, 1)$, then,

$$\hat{V}^{i+1}(x) \leq \hat{V}^i(x) + 2c\check{V}^i(x), \forall x \in \Omega. \quad (81)$$

Proof: Initially note that $\hat{x}_k^{*,i}$ is the k th state vector on the state trajectory initiated from x_0 and propagated by applying control policy $\hat{h}^{(i-1)-k}(\cdot)$ at time k , $0 \leq k \leq i$. The first iteration of AVI leads to

$$\hat{V}^1(x) = \min_{u \in \mathbb{R}^m} \left(U(x, u) + \hat{V}^0(f(x, u)) \right) + \epsilon^0(x), \forall x \in \Omega. \quad (82)$$

One has $\hat{V}^1(x) - \epsilon^0(x) \leq \hat{V}^0(x) - \epsilon^{-1}(x)$, $\forall x$, because, per (82), $\hat{V}^1(x)$ is resulted from a minimization, as opposed to using a given policy $h^{-1}(\cdot)$ in (80). The foregoing inequality along with $-cU(x, 0) \leq -\epsilon^i(x) \leq cU(x, 0)$, $\forall i$, lead to

$$\hat{V}^1(x) \leq \hat{V}^0(x) + 2cU(x, 0), \forall x, \quad (83)$$

which confirms that inequality (81) holds for $i = 0$. Now, assume that

$$\hat{V}^i(x) \leq \hat{V}^{i-1}(x) + 2c\check{V}^{i-1}(x), \forall x \in \Omega, \quad (84)$$

if this assumption leads to (81) the proof will be complete by

induction. Let

$$\hat{\mathcal{V}}(x) := U(x, \hat{h}^{i-1}(x)) + \hat{V}^i(f(x, \hat{h}^{i-1}(x))) + \epsilon^i(x), \forall x \in \Omega. \quad (85)$$

Since the minimizer of the right hand side of (73) is $\hat{h}^i(\cdot)$ and not $\hat{h}^{i-1}(\cdot)$, one has

$$\hat{V}^{i+1}(x) \leq \hat{\mathcal{V}}(x), \forall x \in \Omega. \quad (86)$$

On the other hand, by definition of $\hat{V}^i(x_0)$, that is,

$$\hat{V}^i(x) = U(x, \hat{h}^{i-1}(x)) + \hat{V}^{i-1}(f(x, \hat{h}^{i-1}(x))) + \epsilon^{i-1}(x), \forall x \in \Omega, \quad (87)$$

and comparing it with (85) and considering (84) one has

$$\hat{\mathcal{V}}(x) - \epsilon^i(x) \leq \hat{V}^i(x) - \epsilon^{i-1}(x) + 2c\check{V}^{i-1}(f(x, \hat{h}^{i-1}(x))), \forall x \in \Omega, \quad (88)$$

which because of $-cU(x, 0) \leq -\epsilon^i(x) \leq cU(x, 0)$, $\forall i$, leads to

$$\hat{\mathcal{V}}(x) \leq \hat{V}^i(x) + 2cU(x, 0) + 2c\check{V}^{i-1}(f(x, \hat{h}^{i-1}(x))), \forall x \in \Omega, \quad (89)$$

and along with (86) leads to

$$\hat{V}^{i+1}(x) \leq \hat{V}^i(x) + 2cU(x, 0) + 2c\check{V}^{i-1}(f(x, \hat{h}^{i-1}(x))), \forall x \in \Omega. \quad (90)$$

The next step is showing that

$$U(x_0, 0) + \check{V}^{i-1}(f(x_0, \hat{h}^{i-1}(x_0))) = \check{V}^i(x_0), \forall x_0 \in \Omega. \quad (91)$$

Note that, $\check{V}^i(x_0)$ is the result of evaluating a finite sum of $U(x_k, 0)$'s along a ‘trajectory’ initiated from x_0 . So, in order to show that (91) holds, it suffices to show that the summations in both sides of (91), which each has $i+1$ elements, are along the same trajectory.

The first summand in $\check{V}^i(x_0)$ is $U(x_0, 0)$ which is matched by the same term existing in the left hand side of (91). The second summand of $\check{V}^i(x_0)$ is $U(\cdot, 0)$ evaluated at $\hat{x}_1^{*,i} = f(\hat{x}_0^{*,i}, \hat{h}^{i-1}(\hat{x}_0^{*,i}))$. The first summand of $\hat{V}^{i+1}(f(x_0, \hat{h}^{i-1}(x_0)))$ is $U(\cdot, 0)$ evaluated at $x_1 := f(x_0, \hat{h}^{i-1}(x_0))$. Since $\hat{x}_0^{*,i} = x_0$, one has $\hat{x}_1^{*,i} = x_1$, hence, the second summand of $\check{V}^i(x_0)$ also will be matched by a term in the left hand side of (91). Similarly, the third summand of $\check{V}^i(x_0)$ is $U(\cdot, 0)$ evaluated at $\hat{x}_2^{*,i} = f(\hat{x}_1^{*,i}, \hat{h}^{i-2}(\hat{x}_1^{*,i}))$. The second summand of $\hat{V}^{i+1}(f(x_0, \hat{h}^{i-1}(x_0)))$ is $U(\cdot, 0)$ evaluated at $x_2 := f(x_1, \hat{h}^{i-2}(x_1))$, by definition. Since $\hat{x}_1^{*,i} = x_1$ one has $\hat{x}_2^{*,i} = x_2$. Repeating this argument it is seen that the trajectories are identical and hence, (91) holds, which along with (90) proves (81). \square

Theorem 7. Let the stabilizing approximate value iteration be conducted using a continuous function approximator with the bounded approximation error $|\epsilon^i(x)| \leq cU(x, 0)$, $\forall i \in \mathbb{N} \cup \{-1\}$, $\forall x \in \Omega$, for some $c \in [0, 1)$. Moreover, denoting the value function of $h^{-1}(\cdot)$ with $V^0(\cdot)$, let $\gamma \in \mathbb{R}_+$ be such that $V^0(x) \leq \gamma U(x, 0)$, $\forall x \in \Omega$. Then, for every given $i \in \mathbb{N}$, control policy $\hat{h}^i(\cdot)$ asymptotically stabilizes the system about the origin if c is such that

$$0 \leq c < 1 + 2\gamma - \sqrt{4\gamma^2 + 4\gamma}. \quad (92)$$

Moreover, let the compact domain $\hat{\mathcal{B}}_r^i$ for any $r \in \mathbb{R}_+$ be

defined as $\hat{\mathcal{B}}_r^i := \{x \in \mathbb{R}^n : \hat{V}^i(x) \leq r\}$ and let $\bar{r}^i > 0$ be (the largest r) such that $\hat{\mathcal{B}}_{\bar{r}^i}^i \subset \Omega$. Then, $\hat{\mathcal{B}}_{\bar{r}^i}^i$ will be an estimation of the region of attraction for the system.

Proof: The idea, similar to Theorem 2, is using $\hat{V}^i(\cdot)$ as a Lyapunov function to prove the claim. The lower and upper boundedness of the function, established in Theorem 6, guarantees the positive definiteness of the function and the continuity of the parametric function approximator guarantees its continuity. The objective is showing negative definiteness of $\Delta \hat{V}^i(x) := \hat{V}^i(f(x, \hat{h}^i(x))) - \hat{V}^i(x)$. By Eq. (75)

$$\hat{V}^i(f(x, \hat{h}^i(x))) - \hat{V}^{i+1}(x) = -U(x, \hat{h}^i(x)) - \epsilon^i(x), \forall x \in \Omega. \quad (93)$$

Lemma 6 and inequality (81) may be used in the foregoing equation to replace $\hat{V}^{i+1}(x)$ with $\hat{V}^i(x)$ in its left hand side. Before that, considering the similarity between the exact and approximate VI, from Eqs. (59), (63), (64), and (73), it is straight forward to see

$$\begin{aligned} \hat{V}^i(x_0) &= \hat{V}^0(\hat{x}_k^{*,i}) + \\ &\sum_{k=0}^{i-1} \left(U(\hat{x}_k^{*,i}, \hat{h}^{(i-1)-k}(\hat{x}_k^{*,i})) + \epsilon^{(i-1)-k}(\hat{x}_k^{*,i}) \right), \quad (94) \\ &\forall x_0 \in \Omega, \forall i \in \mathbb{N} - \{0\}, \end{aligned}$$

where $\hat{x}_k^{*,i} := f(\hat{x}_{k-1}^{*,i}, \hat{h}^{i-k}(\hat{x}_{k-1}^{*,i}))$ and $\hat{x}_0^{*,i} := x_0, \forall i, \forall k \in \mathbb{N} - \{0\}$. The point in concluding (94) from the exact VI counterpart, given by (59) considering (63) and (64), is the fact that at time k when the control policy is $\hat{h}^{(i-1)-k}(\cdot)$, the approximation error introduced in the summation will be $\epsilon^{(i-1)-k}(\cdot)$, i.e., the superscripts will be the same per (75) and both functions will be evaluated at the current state $\hat{x}_k^{*,i}$.

Considering $(1-c)U(x, 0) \leq U(x, u) + \epsilon^i(x), \forall x, \forall u, \forall i$, and comparing (94) with $\check{V}(x_0)$ defined in Lemma 6 one has

$$(1-c)\check{V}^i(x) \leq \hat{V}^i(x), \forall x \in \Omega. \quad (95)$$

Moreover, by Theorem 6 one has $\hat{V}^i(x) \leq \bar{V}^i(x), \forall x$, if $\bar{V}^i(\cdot)$ is generated using the value function of $h^{-1}(\cdot)$ as the initial guess. Moreover, $\bar{V}^i(x) \leq \bar{V}^0(x), \forall x$, by Lemma 1. Therefore,

$$\check{V}^i(x) \leq \frac{1}{1-c} \bar{V}^0(x), \forall x \in \Omega, \forall i \in \mathbb{N}. \quad (96)$$

The interesting point about inequality (96) is showing the boundedness of the left hand side. This boundedness along with the results from Lemma 6 lead to

$$\hat{V}^{i+1}(x) \leq \hat{V}^i(x) + \frac{2c}{1-c} \bar{V}^0(x), \forall x \in \Omega, \forall i \in \mathbb{N}. \quad (97)$$

Utilizing (97) in (93) leads to

$$\begin{aligned} \hat{V}^i(f(x, \hat{h}^i(x))) - \hat{V}^i(x) &\leq \\ -U(x, \hat{h}^i(x)) - \epsilon^i(x) &+ \frac{2c}{1-c} \bar{V}^0(x), \forall x \in \Omega. \end{aligned} \quad (98)$$

in order to have $\Delta \hat{V}^i(x) < 0$ one needs $2c/(1-c)\bar{V}^0(x) < U(x, \hat{h}^i(x)) + \epsilon^i(x), \forall x$, which holds if

$$\begin{aligned} \frac{2c}{1-c} \bar{V}^0(x) &< (1-c)U(x, 0) \Leftrightarrow \\ \frac{2c}{(1-c)^2} &< \frac{U(x, 0)}{\bar{V}^0(x)}, \forall x \in \Omega, \end{aligned} \quad (99)$$

Note that $\bar{V}^0(x) \leq 2V^0(x), \forall x$, by definition of $\bar{V}^0(x)$ which is the value function of $h^{-1}(\cdot)$ with the utility of $U(x_k, u_k) + cU(x_k, 0)$, while, $V^0(x)$ is the value function of the same control policy with the utility of $U(x_k, u_k)$. Therefore, from $V^0(x) \leq \gamma U(x, 0)$ one has $\bar{V}^0(x) \leq 2\gamma U(x, 0)$. Hence, $U(x, 0)/\bar{V}^0(x) \geq 1/(2\gamma), \forall x$. Therefore, if $2c/(1-c)^2 < 1/(2\gamma)$ or equivalently if

$$c^2 - (2+4\gamma)c + 1 > 0, \quad (100)$$

then inequality (99) holds. The root of the left hand side of the foregoing inequality are real and given by

$$c_1 = 1 + 2\gamma - \sqrt{4\gamma^2 + 4\gamma} \text{ and } c_2 = 1 + 2\gamma + \sqrt{4\gamma^2 + 4\gamma}. \quad (101)$$

Inequality (100) holds if $c < c_1$ or if $c > c_2$ by analysis of the sign of the quadratic equation on its left hand side. But, $c_2 > 1$, hence, any c which satisfies $c > c_2$ will be unacceptable for our purpose, because, such a c does not belong to $[0, 1]$. As for $c < c_1$ it is required to make sure $c_1 > 0$, otherwise no suitable c will be resulted from this analysis. From $4\gamma^2 + 4\gamma + 1 > 4\gamma^2 + 4\gamma$ which along with the non-negativeness of both sides of the last inequality leads to $\sqrt{4\gamma^2 + 4\gamma + 1} = 2\gamma + 1 > \sqrt{4\gamma^2 + 4\gamma}$, one has $1 + 2\gamma - \sqrt{4\gamma^2 + 4\gamma} > 0$. Therefore, c_1 is indeed positive and a non-negative c smaller than c_1 leads to the desired stability. The first part of the theorem is proved by noticing that when (92) holds, $\Delta \hat{V}^i(x)$ is strictly less than zero, considering the assumed positive definiteness of $U(\cdot, 0)$.

Finally, considering the line of proof at the end of the proof of Theorem 2 it is straight forward to see that $\Delta \hat{V}^i(x) < 0$ leads to $\hat{\mathcal{B}}_{\bar{r}^i}^i \subset \Omega$ being an EROA for the system operated with $\hat{h}^i(\cdot)$. The reason is, any trajectory initiated within $\hat{\mathcal{B}}_{\bar{r}^i}^i$ will remain inside the set and hence, within Ω . Note that the continuity of the function approximators leads to the desired continuity of $\hat{V}^i(\cdot)$ for guaranteeing the compactness of the EROA, as detailed in the proof of Theorem 2. \square

Theorem 8. *Let the stabilizing approximate value iteration be implemented using a continuous function approximator with the bounded approximation error $|\epsilon^i(x)| \leq cU(x, 0), \forall i \in \mathbb{N} \cup \{-1\}, \forall x \in \Omega$, for some $c \in [0, 1]$. Moreover, let $\gamma \in \mathbb{R}_+$ be such that $V^0(x) \leq \gamma U(x, 0), \forall x \in \Omega$. Then, the system operated using control policy $\hat{h}^k(\cdot)$ at time k , that is, the control subject to adaptation in the approximate value iterations scheme asymptotically stabilizes the system about the origin if c is such that*

$$0 \leq c < 1 + 4\gamma - \sqrt{16\gamma^2 + 8\gamma}. \quad (102)$$

Proof: The idea for proof of this theorem is similar to that of Theorem 3, except that the ‘semi-monotonicity’ feature presented in Lemma 6 will be used, instead of the monotonicity given by Lemma 1 in exact VI. Eq. (75) and Lemma 6 lead to

$$\begin{aligned} \hat{V}^1(x_0) &= U(x_0, \hat{h}^0(x_0)) + \hat{V}^0(f(x_0, \hat{h}^0(x_0))) + \\ \epsilon^0(x_0) &\leq \hat{V}^0(x_0) + 2c\check{V}^0(x_0), \forall x_0 \in \Omega, \end{aligned} \quad (103)$$

and similarly

$$\begin{aligned} \hat{V}^2(x) &= U(x, \hat{h}^1(x)) + \hat{V}^1(f(x, \hat{h}^1(x))) + \\ \epsilon^1(x) &\leq \hat{V}^1(x) + 2c\check{V}^1(x), \forall x \in \Omega. \end{aligned} \quad (104)$$

which along with (103) leads to

$$U(x, \hat{h}^1(x)) + \hat{V}^1(f(x, \hat{h}^1(x))) + \epsilon^1(x) \leq \hat{V}^0(x) + 2c\check{V}^0(x) + 2c\check{V}^1(x), \forall x \in \Omega. \quad (105)$$

or equivalently

$$U(x, \hat{h}^1(x)) + \hat{V}^1(f(x, \hat{h}^1(x))) + \epsilon^1(x_0) - 2c\check{V}^0(x) - 2c\check{V}^1(x) \leq \hat{V}^0(x), \forall x \in \Omega. \quad (106)$$

Let $\hat{x}_k^* := f(\hat{x}_{k-1}^*, \hat{h}^{k-1}(\hat{x}_{k-1}^*))$ for $k \geq 1$ and $\hat{x}_0^* := x_0$. Evaluating (106) at \hat{x}_1^* and replacing the $\hat{V}^0(\hat{x}_1^*)$ in the left hand side of the inequality in (103) with the left hand side of (106), which is smaller per (106), one has

$$U(\hat{x}_0^*, \hat{h}^0(\hat{x}_0^*)) + U(\hat{x}_1^*, \hat{h}^1(\hat{x}_1^*)) + \hat{V}^1(\hat{x}_2^*) + \epsilon^0(\hat{x}_0^*) + \epsilon^1(\hat{x}_1^*) - 2c\check{V}^0(\hat{x}_0^*) - 2c\check{V}^0(\hat{x}_1^*) - 2c\check{V}^1(\hat{x}_1^*) \leq \hat{V}^0(\hat{x}_0^*), \forall \hat{x}_0^* \in \Omega. \quad (107)$$

Repeating this process by replacing $\hat{V}^1(\hat{x}_2^*)$ in (107) using

$$U(x, \hat{h}^2(x)) + \hat{V}^2(f(x, \hat{h}^2(x))) + \epsilon^2(x) - 2c\check{V}^1(x) - 2c\check{V}^2(x) \leq \hat{V}^1(x), \forall x \in \Omega, \quad (108)$$

which is resulted from

$$\hat{V}^3(x) = U(x, \hat{h}^2(x)) + \hat{V}^2(f(x, \hat{h}^2(x))) + \epsilon^2(x) \leq \hat{V}^2(x) + 2c\check{V}^2(x) \leq \hat{V}^1(x) + 2c\check{V}^1(x) + 2c\check{V}^2(x), \forall x \in \Omega, \quad (109)$$

leads to

$$U(\hat{x}_0^*, \hat{h}^0(\hat{x}_0^*)) + U(\hat{x}_1^*, \hat{h}^1(\hat{x}_1^*)) + U(\hat{x}_2^*, \hat{h}^2(\hat{x}_2^*)) + \hat{V}^2(\hat{x}_3^*) + \epsilon^0(\hat{x}_0^*) + \epsilon^1(\hat{x}_1^*) + \epsilon^2(\hat{x}_2^*) - 2c\check{V}^0(\hat{x}_0^*) - 2c\check{V}^0(\hat{x}_1^*) - 2c\check{V}^1(\hat{x}_1^*) - 2c\check{V}^1(\hat{x}_2^*) - 2c\check{V}^2(\hat{x}_2^*) \leq \hat{V}^0(\hat{x}_0^*), \forall \hat{x}_0^* \in \Omega. \quad (110)$$

Similarly by repeating this process one has

$$\sum_{k=0}^{i-1} \left(U(\hat{x}_k^*, \hat{h}^k(\hat{x}_k^*)) + \epsilon^k(\hat{x}_k^*) \right) + \hat{V}^{i-1}(\hat{x}_i^*) - 2c \sum_{k=0}^{i-1} \check{V}^k(\hat{x}_k^*) - 2c \sum_{k=1}^{i-1} \check{V}^{k-1}(\hat{x}_k^*) \leq \hat{V}^0(\hat{x}_0^*), \quad \forall \hat{x}_0^* \in \Omega, \forall i \in \mathbb{N} - \{0\}. \quad (111)$$

From (96) and $\bar{V}^0(x) \leq 2V^0(x) \leq 2\gamma U(x, 0), \forall x \in \Omega$, one has

$$\check{V}^i(x) \leq \frac{2\gamma}{1-c} U(x, 0), \forall x \in \Omega, \forall i \in \mathbb{N}. \quad (112)$$

Also, $(1-c)U(x, 0) \leq U(x, u) + \epsilon^i(x), \forall x, \forall u, \forall i$. Therefore, from (111) one has

$$(1-c) \sum_{k=0}^{i-1} U(\hat{x}_k^*, 0) + \hat{V}^{i-1}(\hat{x}_i^*) - \frac{4c\gamma}{1-c} \sum_{k=0}^{i-1} U(\hat{x}_k^*, 0) - \frac{4c\gamma}{1-c} \sum_{k=1}^{i-1} U(\hat{x}_k^*, 0) \leq \hat{V}^0(\hat{x}_0^*), \forall \hat{x}_0^* \in \Omega, \forall i \in \mathbb{N} - \{0\}. \quad (113)$$

The last inequality holds even if we add the negative scalar value of $-4c\gamma/(1-c)U(\hat{x}_0^*, 0)$ to the last summation on its

left hand side, which leads to

$$\left((1-c) - \frac{8c\gamma}{1-c} \right) \sum_{k=0}^{i-1} U(\hat{x}_k^*, 0) + \hat{V}^{i-1}(\hat{x}_i^*) \leq \hat{V}^0(\hat{x}_0^*), \forall \hat{x}_0^* \in \Omega, \forall i \in \mathbb{N} - \{0\}. \quad (114)$$

Since $\hat{V}^{i-1}(x) \geq 0, \forall x$, it can be removed from the left hand side of the foregoing inequality while the inequality still holds. Assume

$$(1-c) - \frac{8c\gamma}{1-c} > 0 \Leftrightarrow c^2 - (2+8\gamma)c + 1 > 0. \quad (115)$$

Comparing (114), after removing $\hat{V}^{i-1}(x_i^*)$, with (57) where the states were propagated using the exact VI in the latter, the same stability result can be obtained for the states propagated using the approximate VI, providing (115) holds. That is, the sequence of partial sums of the left hand side of (114) is upper bounded and because of being non-decreasing it converges, as $i \rightarrow \infty$, [26]. Considering the positive definiteness of $U(\cdot, 0)$, assumed in our AVI analyses, this leads to $x_i \rightarrow 0$, as long as the entire state trajectory is contained in Ω . Finally, in order to enforce (115), one will need (102). The details are identical to the last stages of the proof of Theorem 7, as replacing γ in (100) with 2γ , the left hand sides of inequalities (100) and (115) become identical. Moreover, the proof of positiveness of the right hand side of (102) follows from the same argument presented in that proof. \square

Finally, the last step of our analysis is presenting some results regarding the EROA for the system operated using evolving control policy during AVI.

Theorem 9. *Let the conditions of Theorem 8 hold. Moreover, let $\hat{\mathcal{B}}_r^i := \{x \in \mathbb{R}^n : \hat{V}^i(x) \leq r\}$ and $\mathcal{B}_r^* := \{x \in \mathbb{R}^n : V^*(x) \leq r\}$ for any $r \in \mathbb{R}_+$. If $\mathcal{B}_r^* \subset \Omega$ for a given $r > 0$, then compact set $\hat{\mathcal{B}}_{r/2}^0$ is an estimation of the region of attraction of the system operated using control policy $\hat{h}^k(\cdot)$ at time k , that is, the control subject to adaptation.*

Proof: Once the conditions of Theorem 8 hold, from (114) one has

$$\hat{V}^{i-1}(\hat{x}_i^*) \leq \hat{V}^0(\hat{x}_0^*), \forall \hat{x}_0^* \in \Omega, \forall i \in \mathbb{N} - \{0\}, \quad (116)$$

where $\hat{x}_k^* := f(\hat{x}_{k-1}^*, \hat{h}^{k-1}(\hat{x}_{k-1}^*))$ for $k \geq 1$ and $\hat{x}_0^* := x_0$, that is, \hat{x}_k^* denotes the state trajectory generated by the evolving control policy from the AVI. Therefore,

$$x_0 \in \hat{\mathcal{B}}_r^0 \Rightarrow \hat{x}_k^* \in \hat{\mathcal{B}}_r^{k-1}, \forall k \in \mathbb{N} - \{0\}, \forall r \in \mathbb{R}_+. \quad (117)$$

From $\underline{V}^k(x) \leq \hat{V}^k(x), \forall x, \forall k$, cf. Theorem 6, one has

$$\hat{\mathcal{B}}_r^k \subset \underline{\mathcal{B}}_r^k, \forall k \in \mathbb{N}, \forall r \in \mathbb{R}_+, \quad (118)$$

where $\underline{\mathcal{B}}_r^i := \{x \in \mathbb{R}^n : \underline{V}^i(x) \leq r\}$ and $\underline{V}^k(x)$ is defined in Theorem 6. Moreover, $\{\underline{V}^k(x)\}_{k=0}^\infty$ is non-increasing and converges to $\underline{V}^*(x)$, i.e., the optimal value function corresponding to cost function (76), because it is resulted from an exact VI. Therefore, $\underline{V}^*(x) \leq \underline{V}^{k+1}(x) \leq \underline{V}^k(x), \forall x$, which leads to

$$\underline{\mathcal{B}}_r^k \subset \underline{\mathcal{B}}_r^{k+1} \subset \underline{\mathcal{B}}_r^*, \forall k \in \mathbb{N}, \forall r \in \mathbb{R}_+. \quad (119)$$

From (118) and (119) one has

$$\hat{\mathcal{B}}_r^k \subset \underline{\mathcal{B}}_r^*, \forall k \in \mathbb{N}, \forall r \in \mathbb{R}_+. \quad (120)$$

On the other hand, it is not hard to see that $V^*(x) \leq$

$2V^*(x), \forall x$, which if holds leads to

$$\mathcal{B}_r^* \subset \mathcal{B}_{2r}^*, \forall r \in \mathbb{R}_+. \quad (121)$$

To verify $V^*(x) \leq 2V^*(x)$, let the state trajectory generated from the exact VI of $V^*(x)$ be denoted with \underline{x}_k^* , $\forall k$. Comparing cost function (76) with (2) one has

$$V^*(x_0) \leq V^*(x_0) + c \sum_{k=0}^{\infty} U(\underline{x}_k^*, 0), \forall x_0. \quad (122)$$

The reason is, both sides of the foregoing inequality are infinite sums of similar summands, except that the the left hand side is evaluated along the optimal trajectory with respect to (2) while the right hand side is evaluated along the optimal trajectory with respect to (76). Moreover, by composition $\sum_{k=0}^{\infty} U(\underline{x}_k^*, 0) \leq V^*(x_0)$ which along with $c < 1$ and (122) leads to $V^*(x) \leq 2V^*(x)$.

Finally, from (117), (120), and (121) one has

$$x_0 \in \hat{\mathcal{B}}_r^0 \Rightarrow \hat{x}_k^* \in \mathcal{B}_{2r}^*, \forall k \in \mathbb{N} - \{0\}, \forall r \in \mathbb{R}_+. \quad (123)$$

Therefore, any state trajectory initiated within $\hat{\mathcal{B}}_{r/2}^0$ remains inside \mathcal{B}_r^* and if the r in the latter is such that $\mathcal{B}_r^* \subset \Omega$, the state trajectory remains within Ω and by Theorem 8 converges to the origin. \square

VI. CONCLUSIONS

A set of theoretical analyses on convergence, stability, and regions of attraction for a value iteration based scheme which is initiated using an admissible guess were presented. Afterwards, the results were extended to the more interesting but challenging case of incorporating the approximation errors in the iterations. Simple and straight forward conditions for guaranteeing the boundedness of the learning results and the stability of the system under a fixed as well as an evolving control policy were developed. These results are expected to lay the foundation for improving the mathematical rigor of the popular field of intelligent control.

REFERENCES

- [1] C. Watkins, *Learning from Delayed Rewards*. PhD Dissertation, Cambridge University, Cambridge, England, 1989.
- [2] P. J. Werbos, "Approximate dynamic programming for real-time control and neural modeling," in *Handbook of Intelligent Control* (D. A. White and D. A. Sofge, eds.), Multiscience Press, 1992.
- [3] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. MIT Press, 1998.
- [4] D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-Dynamic Programming*. Athena Scientific, 1996.
- [5] W. B. Powell, *Approximate Dynamic Programming*. Hoboken, NJ, Wiley, 2007.
- [6] S. N. Balakrishnan and V. Biega, "Adaptive-critic based neural networks for aircraft optimal control," *Journal of Guidance, Control and Dynamics*, vol. 19, pp. 893–898, 1996.
- [7] D. Prokhorov and D. Wunsch, "Adaptive critic designs," *IEEE Transactions on Neural Networks*, vol. 8, pp. 997–1007, 1997.
- [8] G. Venayagamoorthy, R. Harley, and D. Wunsch, "Comparison of heuristic dynamic programming and dual heuristic programming adaptive critics for neurocontrol of a turbogenerator," *IEEE Transactions on Neural Networks*, vol. 13, pp. 764–773, May 2002.
- [9] R. Enns and J. Si, "Helicopter trimming and tracking control using direct neural dynamic programming," *IEEE Transactions on Neural Networks*, vol. 14, no. 4, pp. 929–939, 2003.
- [10] A. Al-Tamimi, F. Lewis, and M. Abu-Khalaf, "Discrete-time nonlinear HJB solution using approximate dynamic programming: Convergence proof," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 38, pp. 943–949, Aug 2008.

- [11] Q. Zhao, H. Xu, and S. Jagannathan, "Optimal control of uncertain quantized linear discrete-time systems," *International Journal of Adaptive Control and Signal Processing*, 2014.
- [12] D. Liu and Q. Wei, "Finite-approximation-error-based optimal control approach for discrete-time nonlinear systems," *IEEE Transactions on Cybernetics*, vol. 43, pp. 779–789, April 2013.
- [13] A. Heydari and S. N. Balakrishnan, "Fixed-final-time optimal control of nonlinear systems with terminal constraints," *Neural Networks*, vol. 48, pp. 61–71, 2013.
- [14] F. Lewis, D. Vrabie, and K. Vamvoudakis, "Reinforcement learning and feedback control: Using natural decision methods to design optimal adaptive controllers," *IEEE Control Systems*, vol. 32, pp. 76–105, 2012.
- [15] D. Liu and Q. Wei, "Policy iteration adaptive dynamic programming algorithm for discrete-time nonlinear systems," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, pp. 621–634, 2014.
- [16] T. Landelius, *Reinforcement learning and distributed local model synthesis*. PhD Dissertation, Linköping Univ., Linköping, Sweden, 1997.
- [17] X. Liu and S. Balakrishnan, "Convergence analysis of adaptive critic based optimal control," in *Proceedings of the American Control Conference*, vol. 3, pp. 1929–1933, 2000.
- [18] B. Lincoln and A. Rantzer, "Relaxing dynamic programming," *IEEE Transactions on Automatic Control*, vol. 51, pp. 1249–1260, Aug 2006.
- [19] M. Rinehart, M. Dahleh, and I. Kolmanovsky, "Value iteration for (switched) homogeneous systems," *IEEE Transactions on Automatic Control*, vol. 54, no. 6, pp. 1290–1294, 2009.
- [20] A. Heydari, "Revisiting approximate dynamic programming and its convergence," *IEEE Transactions on Cybernetics*, vol. 44, no. 12, pp. 2733–2743, 2014.
- [21] S. P. Singh and R. C. Yee, "An upper bound on the loss from approximate optimal-value functions," *Mach. Learn.*, vol. 16, pp. 227–233, 1994.
- [22] A. Farahmand, C. Szepesvári, and R. Munos, "Error propagation for approximate policy and value iteration," in *Advances in Neural Information Processing Systems* (J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, eds.), pp. 568–576, 2010.
- [23] R. Munos and C. Szepesvári, "Finite-time bounds for fitted value iteration," *J. Mach. Learn. Res.*, vol. 9, pp. 815–857, 2008.
- [24] Q. Wei and D. Liu, "A novel iterative θ -adaptive dynamic programming for discrete-time nonlinear systems," *IEEE Transactions on Automation Science and Engineering*, vol. 11, no. 4, pp. 1176–1190, 2014.
- [25] H. Khalil, *Nonlinear Systems*. Prentice-Hall, 2002. pp. 111–194.
- [26] W. Rudin, *Principles of Mathematical Analysis*. McGraw-Hill, 3rd ed., 1976. pp. 55, 60, 86, 87, 89, 145, 148.
- [27] A. Heydari, "Theoretical and Numerical Analysis of Approximate Dynamic Programming with Approximation Errors," available at arXiv.org.
- [28] H. Jeffreys and B. S. Jeffreys, "Weierstrass's theorem on approximation by polynomials," in *Methods of Mathematical Physics*, pp. 446–448, Cambridge University Press, 3rd ed., 1988.
- [29] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Networks*, vol. 2, no. 5, pp. 359–366, 1989.
- [30] D. E. Kirk, *Optimal control theory; an introduction*. Prentice-Hall, 1970. pp. 53–106.
- [31] A. Heydari and S. Balakrishnan, "Optimal switching between autonomous subsystems," *Journal of the Franklin Institute*, vol. 351, 2014.
- [32] A. Rantzer, "Relaxed dynamic programming in switching systems," *Control Theory and Applications, IEE Proceedings*, vol. 153, no. 5, pp. 567–574, 2006.
- [33] A. Jadbabaie, J. Yu, and J. Hauser, "Unconstrained receding-horizon control of nonlinear systems," *IEEE Transactions on Automatic Control*, vol. 46, no. 5, pp. 776–783, 2001.
- [34] E. Sontag, "A 'universal' construction of Artstein's theorem on nonlinear stabilization," *System & Control Letters*, vol. 13, pp. 117–123, 1989.



Ali Heydari received his PhD degree from the Missouri University of Science and Technology in 2013. He is currently an assistant professor of mechanical engineering at the South Dakota School of Mines and Technology. He was the recipient of the Outstanding M.Sc. Thesis Award from the Iranian Aerospace Society, the Best Student Paper Runner-Up Award from the AIAA Guidance, Navigation and Control Conference, and the Outstanding Graduate Teaching Award from the Academy of Mechanical and Aerospace Engineers at Missouri S&T. His research interests include approximate dynamic programming and control of hybrid systems. He is a member of Tau Beta Pi.